Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

JANNES QUER [1] AND HAN LIE [2]

Some connections between importance sampling and enhanced sampling methods in molecular dynamics

[1]Zuse Institute Berlin
[2]Zuse Institute Berlin and Freie Universität Berlin Mathematics Department

ZIB Report 17-30 (Juni 2017)

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30-84185-0
Telefax: +49 30-84185-125

E-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Some connections between importance sampling and enhanced sampling methods in molecular dynamics

Jannes Quer and Han Lie

June 9, 2017

Abstract

Enhanced sampling methods play an important role in molecular dynamics, because they enable the collection of better statistics of rare events that are important in many physical phenomena. We show that many enhanced sampling methods can be viewed as methods for performing importance sampling, by identifying important correspondences between the language of molecular dynamics and the language of probability theory. We illustrate these connections by highlighting the similarities between the rare event simulation method of Hartmann and Schütte (*J. Stat. Mech. Theor. Exp.*, 2012), and the enhanced sampling method of Valsson and Parrinello (*Phys. Rev. Lett.* 113, 090601). We show that the idea of changing a probability measure is fundamental to both enhanced sampling and importance sampling.

1 Introduction

One of the main limitations of Molecular Dynamics (MD) simulations is the so called sampling problem [Bernardi et al., 2015]. This is why many enhanced sampling techniques, e.g. umbrella sampling [Torrie and Valleau, 1977], Metadynamics [Laio and Parrinello, 2002], replica exchange (parallel tempering) [Swendsen and Wang, 1986], and simulated annealing [Kirkpatrick et al., 1983] have been developed. The mathematical framework of importance sampling underlies many of these techniques. Importance sampling is used in many different fields, from financial mathematics to MD. A deeper understanding of the common features of these different fields can lead to a deeper understanding of the fundamental problems; it can potentially lead to the development of useful strategies for MD, by adapting successful methods that have been applied in other fields. For this reason, we think it is important to point out the fundamental connection between enhanced sampling and importance sampling. The aim of this article is to explicitly formulate enhanced sampling as a method for importance sampling, by identifying correspondences between them. We illustrate these correspondences using two enhanced sampling techniques for complex systems [Valsson and Parrinello, 2014, Hartmann and Schütte, 2012]. We will also comment briefly on other methods which also fit in this mathematical framework.

In MD, one is often interested in the stationary distribution (Boltzmann Gibbs distribution) of the molecular system or different dynamical quantities of interest, such as the transition probabilities between conformations or exit rates. These quantities can be formulated in terms of expectations of path functionals, where the expectation is taken over a trajectory $X_{0:\tau}$ of finite length τ of the molecular system of interest. Since it is in practice impossible to calculate these expectations analytically, one uses a Monte Carlo approximation, e.g. the empirical mean

$$\int_{\Omega} f(X_{0:\tau}) d\mathbb{P} \approx \frac{1}{N} \sum_{i=1}^{N} f(X^i_{0:\tau}), \tag{1}$$

where $X^i_{0:\tau}$ denotes an independent, identically distributed copy of the (random) trajectory $X_{0:\tau}$ of the system, and $\mathbb{P}$ is the measure on path space associated to the equilibrium Boltzmann Gibbs measure. In the simulation and MD communities, these expectations are often called 'ensemble averages'.

Often the systems under investigation are metastable, which implies that the path functional of interest itself is associated to some rare event, e.g. the first time to transition from one metastable conformation to another. The presence of metastability or kinetic bottlenecks ensures that it is difficult to collect good statistics of such path functionals, because, with high probability, the rare event does not occur. The methods mentioned in the first paragraph were developed to overcome this problem, and can be divided into two groups: those which modify the stationary distribution of the molecular system by changing the potential, the force field, or the temperature, e.g. [Laio and Parrinello, 2002, Huber et al., 1994, Wang and Landau, 2001, Valsson and Parrinello, 2014] [Kirkpatrick et al., 1983][Sorensen and Voter, 2000] [Voter, 1997], and those which do not modify the stationary distribution, e.g. [Elber and West, 2010] [Cerou and Guyader, 2006]. The methods that modify the stationary distribution can be viewed as methods for importance sampling, and they are the methods that we shall consider in this article. Since it is impossible to cover all the methods that have been developed for MD, we refer the interested reader to [Leimkuhler and Matthews, 2015, Tuckerman, 2010, Chipot and Pohorille, 2007] [Leliévre et al., 2010] and the references therein for more details.

The article is organized as follows. We motivate the formulation of molecular dynamics in terms of random dynamical systems in Section 2. In Section 3, we present the fundamental ideas of probability theory, Monte Carlo methods, and importance sampling. We show how enhanced sampling and importance sampling are connected in Section 4, using the variational enhanced sampling method of Valsson and Parrinello [Valsson and Parrinello, 2014] and the efficient rare event simulation method of Hartmann and Schütte [Hartmann and Schütte, 2012] as examples. In Section 4.5, We discuss briefly how one can view other enhanced sampling methods, such as the adaptive biasing force method as methods for importance sampling. We comment on the connections to other topics such as stochastic optimal control, machine learning and Bayesian inference, in Section 5.

2 Motivation for random models of molecular dynamics

One of the most common justifications for modeling a physical phenomenon of interest using random variables is the often complicated dependence of the phenomenon on a large number of factors. In molecular dynamics, one often wishes to study the dynamical behavior of a large molecule immersed in a bath of a very large number of solvent molecules. Following Newton's laws, the molecule evolves according to the sum of forces acting on it. These forces belong to one of two categories: 1) the forces arising from the bonded and non-bonded (e.g. electrostatic) interactions involving the constituent atoms; and 2) the forces arising from the interactions of the molecule with surrounding molecules, e.g. solvent molecules. The forces in the first category are related to the potential energy function U which depends on the molecule itself; they are captured by the force field, and lead to the Boltzmann distribution $\exp(-\beta U(x))/Z$ where $\beta^{-1} = k_B T$ is the Boltzmann constant multiplied by the temperature T and x denotes the state or configuration of the molecule. The forces in the second category are often modeled using explicit water models or Brownian dynamics simulations; although their magnitude is on average (much) smaller than the magnitude of the forces from the first category, the forces from the second category are the key to unlocking interesting phenomena, especially when the molecule exhibits metastable dynamics. This is because, given an initial condition, and in the absence of the forces from the second category, the molecule would evolve according to the forces from the first category in order to reach the nearest local minimum of the energy function U. The forces from the second category ensure that the molecule does not come to rest at the local minimum of U nearest to the given initial condition.

In many molecular dynamics simulations, the ratio of solvent molecules to solute molecules is typically several orders of magnitude, so that it is computationally impossible to model the effect of each individual collision of a solvent molecule on the solute molecule; one therefore resorts to a random variable to model the cumulative effect of solvent-solute collisions. Thus, according to this probabilistic model, one can consider a molecular dynamics simulation as a series of random experiments: at every time point in the simulation, one computes the dominant, deterministic forces acting on the molecule arising from the energy function U, as well as the random forces that arise from solvent-solute collisions, and then updates the state of the molecule using the sum of these forces. The resulting sequence of states of the molecule thus form a particular realization of a sequence of random variables, i.e. of a *stochastic process*, and repeating the molecular dynamics simulations yields different realizations. Note that, even when considering deterministic dynamics, the chaotic behavior of molecular systems implies that no two trajectories with the same starting conditions will agree [Leimkuhler and Matthews, 2015, pp. 41].

One stochastic model for molecular dynamics is the overdamped Langevin equation, which is also known as 'Brownian Dynamics'. In mathematical terms, this is a diffusion process given by the stochastic differential equation (SDE) as a model of the atomistic movement of a molecule. This SDE satisfies

$$dX_t = -\nabla U(X_t)dt + \sqrt{2\beta^{-1}}dB_t, \quad X_0 = x. \tag{2}$$

Above, X_t denotes the state of the system at time $t \geq 0$, $\beta > 0$ is a scaling factor for the noise associated with the temperature and the Boltzmann constant often called the inverse temperature and B_t is a standard n-dimensional Brownian motion with respect to the probability measure $\mathbb{P}$ on some probability space $(\Omega, \mathbb{P}, \mathcal{F})$, and $U : \mathbb{R}^n \to \mathbb{R}$ is a sufficiently smooth (e.g. $\mathcal{C}^1$) potential energy function. This model can be extended by including a friction term c.f. [Leimkuhler and Matthews, 2015]. In the stochastics community, the object integrated with respect to dt is called the 'drift', and the object integrated with respect to dB_t term is called the 'diffusion'. One often also assumes that the potential U satisfies growth conditions at infinity, so that $\exp(-\beta U(x))$ is integrable with respect to Lebesgue measure on the finite-dimensional space $\mathbb{R}^n$. Another model is given by the Langevin equation, in which a second equation for the momentum is included, and the stochastic perturbation only acts on the momentum equation. In this article, we focus on the simple overdamped Langevin equation (2), and refer the interested reader to [Leimkuhler and Matthews, 2015] for more detail on the Langevin equation.

3 Importance Sampling and Change of Measure

3.1 Fundamentals of probability theory

Importance sampling is a method for improving the estimation by Monte Carlo of statistical quantities, such as expected values of random variables [Bucklew, 2004]. In order to properly formulate importance sampling, we first define a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all random variables of interest are defined. In the mathematics and statistics literature, Ω is sometimes referred to as the 'event space', while $\mathcal{F}$ is a collection of subsets of Ω with specific properties (in the mathematical terminology, $\mathcal{F}$ is a σ-algebra on Ω). The significance of $\mathcal{F}$ is that it consists of all events for which one can compute fundamental statistical quantities, e.g., probabilities, means, and variances. The measure $\mathbb{P}$ in turn is a function defined on $\mathcal{F}$ that assigns to every element $A \in \mathcal{F}$ a value in the unit interval $[0, 1]$; this value is the probability of the event A. A real-valued random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $X : \Omega \to \mathbb{R}$ such that, given a Borel subset A of $\mathbb{R}$, the event $\{X \in A\}$ that X takes values in A belongs to the σ-algebra $\mathcal{F}$. The concept of a probability space is important because it justifies all the mathematical operations performed in statistical experiments, e.g. scalar multiplication, addition, powers, logarithms, and exponentials.

A particularly nice feature of probability spaces is that they behave well under reasonable transformations: given a nonempty set E and a σ-algebra $\mathcal{E}$ on E, a random variable $X : \Omega \to E$ on $(\Omega, \mathcal{F}, \mathbb{P})$ defines another probability space $(E, \mathcal{E}, \mathbb{P} \circ X^{-1})$, where the probability measure $\mathbb{P} \circ X^{-1}$ on E is known as the *law* or *distribution* of X. The law of X is particularly useful, because it yields a so-called 'change of variables formula'. For example, when X is a real-valued random variable (i.e. when $E = \mathbb{R}$ and $\mathcal{E}$ is the Borel σ-algebra $\mathcal{B}(\mathbb{R})$), we have the following identity for the mean or expectation of X with respect to $\mathbb{P}$,

$$\mathbb{E}[X] := \int_\Omega X(\omega) d\mathbb{P}(\omega) = \int_{\mathbb{R}} x \; d\mathbb{P} \circ X^{-1}(x), \tag{3}$$

where the second equation is precisely the change of variables formula. By defining $F(x) := \mathbb{P} \circ X^{-1}((-\infty, x]) = \mathbb{P}(X \leq x)$, we obtain the cumulative distribution function of X from its law.

3.2 Monte Carlo and importance sampling

Given a real-valued random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, consider the task of estimating the mean of X. For simplicity, we consider only the one-dimensional case. The basic Monte Carlo estimator of $\mathbb{E}[X]$ is the random variable (also known as the *empirical mean*)

$$\widehat{\varrho} = \frac{1}{N} \sum_{i=1}^{N} X^{(i)}$$

where for every $1 \leq i \leq N$, $X^{(i)}$ is an independent, identically distributed (i.i.d.) copy of the random variable of interest X. A consequence of being an identically distributed copy of X is that $\mathbb{E}[X^{(i)}] = \mathbb{E}[X]$ for all $1 \leq i \leq N$. Thus the estimator $\widehat{\varrho}$ is *unbiased*, in the sense that $\mathbb{E}[X] = \mathbb{E}[\widehat{\varrho}]$. If we wish to estimate the mean of a given *function* of X, then we may similarly use the empirical mean estimator

$$\widehat{\varrho}_f := \frac{1}{N} \sum_{i=1}^{N} f(X^{(i)}), \tag{4}$$

which is also an unbiased estimator under the same conditions in the sense that $\mathbb{E}[f(X)] = \mathbb{E}[\widehat{\varrho}_f]$.

We now make an important observation. If

$$\sum_{i=1}^{N} a_i = 1 \text{ and } \mathbb{E}[f(X^{(i)})] = \mathbb{E}[f(X)] \quad \forall i = 1, \ldots, N, \tag{5}$$

then any random variable of the form $\Sigma_{i=1}^{N} a_i f(X^{(i)})$ will be an unbiased estimator of $\mathbb{E}[f(X)]$. In other words, we do not need either independence of the $X^{(i)}$ or that the $X^{(i)}$ are identically distributed copies of X. Given an expected value that we wish to compute, we may construct estimators of this quantity using any random variables, provided that at the very least the conditions in (5) hold. These random variables can be chosen according to user-defined criteria, e.g. the estimator $\widehat{\varrho}$ has a smaller variance than X. Indeed, variance reduction is perhaps the primary consideration, because the smaller the variance of the unbiased estimator $\widehat{\varrho}$, the faster it converges (in some sense) to the quantity of interest as the sample size N grows to infinity, by the law of large numbers.

An important first step in importance sampling is to find a random variable Y that is defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as the random variable X, has the same range E as X, and is distinct from X. These constraints imply that the law of Y and the law of X are two different probability measures on the same pair $(E, \mathcal{E})$. It is important that both X and Y are functions from Ω to E, since we may then consider the property of absolute continuity of laws. The measure $\mathbb{P} \circ X^{-1}$ is *absolutely continuous* with respect to the measure $\mathbb{P} \circ Y^{-1}$ if, for any set $A \in \mathcal{E}$ for which $\mathbb{P} \circ Y^{-1}(A) = 0$, it also holds that $\mathbb{P} \circ X^{-1}(A) = 0$;

one writes $\mathbb{P} \circ X^{-1} \ll \mathbb{P} \circ Y^{-1}$. Absolute continuity can be understood as the property that any event that is statistically impossible according to the law of Y must also be statistically impossible according to the law of X. Two probability measures μ and ν on the same pair $(E, \mathcal{E})$ are said to be *equivalent* or *mutually absolutely continuous* if $\mu \ll \nu$ and $\nu \ll \mu$. The significance of two measures being equivalent is that they share the same collection of events that are statistically impossible. From now on, we shall write $\mu = \mathbb{P} \circ X^{-1}$ for the reference distribution with respect to which our desired statistical quantities are defined, and $\nu = \mathbb{P} \circ Y^{-1}$ for the auxiliary (importance sampling) distribution.

If $\mu \ll \nu$, then there exists a nonnegative function $d\mu/d\nu$ defined on E such that for for any μ-integrable function f on E,

$$\mathbb{E}_\mu[f] \equiv \mathbb{E}[f(X)] = \mathbb{E}\left[f(Y)\frac{d\mu}{d\nu}(Y)\right] \equiv \mathbb{E}_\nu\left[f\frac{d\mu}{d\nu}\right]. \tag{6}$$

The function $d\mu/d\nu$ is called the *Radon-Nikodym derivative* of μ with respect to ν. We emphasize that in general, the Radon-Nikodym derivative is not the ratio of two functions.

The idea of importance sampling is to find, given a random variable X and the desired quantity of interest $\mathbb{E}[f(X)]$, another random variable Y such that the Radon-Nikodym derivative can be evaluated. One often demands Y be such that the empirical mean of $f(Y)\frac{d\mu}{d\nu}(Y)$ using N i.i.d. draws from the law of Y satisfies some user-defined constraints, such as smaller variance.

Note that the approach above proceeded by assuming the existence of a random variable Y such that the law of X was absolutely continuous with respect to the law of Y, and arrived at the existence of the associated Radon-Nikodym derivative. One can proceed in the reverse direction: given the E-valued random variable X, its associated law $\mu = \mathbb{P} \circ X^{-1}$, and a strictly positive function h on E such that $\int_E h(z)d\mu(z)$ is finite, it holds that the normalized function

$$\frac{h(z)}{\int_E h(z')d\mu(z')}$$

is the Radon-Nikodym derivative $d\nu/d\mu$ of some probability measure ν on $(E, \mathcal{E})$, and hence that ν is absolutely continuous with respect to μ. Since every random variable is uniquely determined by its law, and since every law is a probability measure, it follows that there exists a unique random variable whose law is absolutely continuous with respect to the law μ.

3.3 Translating IS into MD language

In order to understand the application of importance sampling in the MD context, we now establish some correspondences between the terminology of importance sampling and that of MD.

In MD, one is often interested in sampling trajectories of a N-atom molecule in its configuration space $\mathbb{R}^{3N}$. This implies that the random variable of interest is a stochastic process; in this case, it is natural to define the event space Ω to be the set of continuous trajectories in the configuration space (we omit the discussion of the σ-algebra $\mathcal{F}$, since it is not important here). For example, if one adopts a random model of the molecular dynamics as in Section 2, then the fundamental random variable of interest is the stochastic process $X_{0:\tau} = (X_t)_{0 \le t \le \tau}$

defined by (2), where $\tau > 0$ is some predefined duration of the trajectory. To simplify notation, we shall omit the duration subscript and write X instead of $X_{0:\tau}$ for the remainder of this article. The stochastic process is determined by the potential function U (whose gradient corresponds to the force field used in the MD simulation), and by the reference measure $\mathbb{P}$ that defines the Brownian motion process $B = (B_t)_{0 \leq t \leq \tau}$. In particular, if either the potential U or the reference measure $\mathbb{P}$ are changed, then the random variable X will also change.

Suppose we perform multiple molecular dynamics simulations of duration $\tau > 0$, each starting from a common initial state x_0, and each using the force field associated to the potential U as given in (2). The law of X on $(\Omega, \mathcal{F})$ then has some intuitive properties. For an arbitrary $0 < t < \tau$, consider the evaluation map $f(X) := X(t)$. Then the function f defines a vector-valued path functional, since it assigns to every path X the vector in $\mathbb{R}^{3N}$ that describes the configuration of the molecule at time t. If the forces acting on the molecule do not change over time, and if the molecule is in equilibrium, then the law of the random variable $f(X)$ will be given by the Boltzmann distribution $\exp(-\beta U(z))/Z(\beta)$. We can consider another example in the same setting. Suppose one is interested in computing the first mean passage time for a conformational change from A to B, where necessarily the initial state x_0 belongs to the conformation A. In this case, we may define the function f to be the first time that the molecule enters the conformation B (provided that the molecule enters the conformation B within the time interval $0 \leq t \leq \tau$), and estimate the mean first passage time from $x_0 \in A$ to B using the empirical mean estimator given in (4). Note that the first passage time depends on the dynamics of the molecule over the entire trajectory. Such quantities are sometimes called *dynamical quantities*, to distinguish them from quantities that only depend on a finite number of time points along the trajectory. In [Hartmann and Schütte, 2012], it was shown that to any dynamical quantity of the form

$$W(X) = \int_0^\tau f(X_s) ds.$$

we may associate a Free Energy Surface (FES) conditioned on the initial condition $X_0 = x$, by

$$F(x) = -\beta \log \mathbb{E}[\exp^{-W(X)/\beta} | X_0 = x]. \tag{7}$$

For example, if one sets f in the definition of W above to be the indicator function of a set $\mathcal{S}$, then the resulting quantity $F(x)$ is related to the cumulant generating function of the occupation time of $\mathcal{S}$.

It is known that, if the molecule is metastable, then transitions between conformations are rare. When these transitions are the phenomena of interest, one often wishes to collect better statistics, i.e. to sample transition events more easily. One way to achieve this would be to add a biasing potential U_{bias} to the energy function U. Recall that a random variable X is uniquely determined by its law $\mu = \mathbb{P} \circ X^{-1}$, and recall that μ is defined by the fact that the distribution of the vector-valued random variable $X(t) \in \mathbb{R}^{3N}$ follows the Boltzmann distribution for every $0 < t \leq \tau$. If we add a biasing potential U_{bias} to the energy function U that defines the Boltzmann distribution, then we *change the measure* from μ to some other measure ν. Therefore, we no longer work with the random variable X, but some other random variable $Y = Y_{0:\tau}$. As in importance

sampling, the hope is then that the modified random variable is more favorable with respect to certain user-defined criteria, e.g. transitions from A to B occur more frequently. If μ and ν are equivalent, i.e. mutually absolutely continuous, then the modified 'molecule' whose random trajectories are described by Y cannot exhibit behavior which is impossible for the original molecule whose trajectories are described by X. The associated Radon-Nikodym derivative can be interpreted in this case as a *reweighting factor* with which we can use the statistics of Y in order to obtain unbiased estimators of statistics of X, using (6). In [Vanden-Eijden and Weare, 2012], it is shown that one can also apply importance sampling techniques to quantities of a similar form to (7) in order to estimate rare event probabilities.

3.4 Different Estimators

In this paragraph we shall show how different Monte Carlo estimators can be constructed. Recall that the random variables $X = X_{0:\tau}$ and $Y = Y_{0:\tau}$ refer to trajectories of duration τ.

The simplest Monte Carlo estimator of the quantity $\rho = \mathbb{E}[f(X)]$ is given by the empirical mean defined in (4).

As discussed in the preceding section, in some circumstances it may be desirable to sample trajectories from an auxiliary (importance sampling) distribution, e.g. when an event of interest occurs with very small probability with respect to the reference distribution, and then re-weight to obtain the statistics according to the reference distribution. Using the notation introduced earlier of the reference distribution $\mu = \mathbb{P} \circ X^{-1}$ and auxiliary distribution $\nu = \mathbb{P} \circ Y^{-1}$, we can define the associated importance sampling estimator by

$$\hat{\varrho} = \frac{1}{N} \sum_{i=1}^{N} f(Y^{(i)}) \frac{d\mu}{d\nu}(Y^{(i)})$$

where $Y^{(i)} = Y^{(i)}_{0:\tau}$ is the i-th sample trajectory drawn from ν, and ν is defined by a Boltzmann Gibbs distribution that differs from that of μ. Recall from Section 3 that the Radon-Nikodym derivative is defined, provided that μ is absolutely continuous with respect to ν. The unbiasedness of the importance sampling estimator was shown in (6).

4 Enhanced Sampling and Importance Sampling

4.1 Comparing the Valsson-Parrinello and Hartmann-Schütte methods

In this section, we explain that enhanced sampling and importance sampling are connected by the idea of change of measure, using two recently published methods [Valsson and Parrinello, 2014, Hartmann and Schütte, 2012] as examples.

The Valsson-Parrinello method described in [Valsson and Parrinello, 2014] involves a variational approach to enhanced sampling for complex systems in collective variables (CVs) or reaction coordinates. Suppose that the CV space is a Euclidean space equipped with Lebesgue measure dx, and that the reference

measure μ (in this case, μ is not a path measure) is associated to some reference potential U. Suppose that one wishes to sample from some prescribed auxiliary distribution ν that has a density h with respect to the Lebesgue measure, i.e. $d\nu(x) = h(x)dx$; an example given in [Valsson and Parrinello, 2014] concerns the uniform distribution on a bounded subset of CV space. Then, by defining the following functional of the biased potential

$$\phi(V) = \frac{1}{\beta} \log \frac{\int e^{-\beta(U(s)+V(s))} ds}{\int e^{-\beta U(s)} ds} + \int h(s)V(s)ds, \tag{8}$$

one can show that the following biasing potential

$$V(s) = -U(s) - \frac{1}{\beta} \log h(s) \tag{9}$$

extremises the functional ϕ. In particular, one can define an optimization problem, the solution of which is given by V. The authors use the fact that the functional F is midpoint convex in order to find the optimal bias, by parametrizing the bias potential and using an optimization scheme to find the optimal parameters. The parametrization is done by an expansion of V into a linear set of basis functions, and the optimization is done by stochastic gradient descent. From the two equations above, it follows that if the target distribution h is strictly positive everywhere, and if $e^{-\beta U}$ is integrable (which implies that U is finite almost everywhere), then μ and ν are mutually equivalent, with

$$\frac{d\mu}{d\nu}(x) = \frac{Z_\mu}{Z_\nu} \frac{e^{-\beta 2U(x)}}{h(x)} \tag{10}$$

where $Z_\mu = \int e^{-\beta U(x)} dx$ is the normalization constant and Z_ν is the normalization constant for the Boltzmann Gibbs measure with respect to the biased potential V. The Hartmann-Schütte method described in [Hartmann and Schütte, 2012] involves a method of performing importance sampling for systems described by overdamped Langevin dynamics for trajectories for $0 \leq t \leq \tau$:

$$dX_t = -\nabla U(X_t)dt + \sqrt{2\beta^{-1}} dB_t, \qquad X_0 = x \tag{11}$$

$$dY_t = -\nabla (V + U)(Y_t)dt + \sqrt{2\beta^{-1}} dB_t, \qquad Y_0 = x. \tag{12}$$

The first SDE is the original dynamics corresponding to the Boltzmann measure $e^{-\beta U}$ and the second SDE is the perturbed dynamics corresponding to the Boltzmann measure $e^{-\beta(U+V)}$; see Section 3.3 for more details. For a given choice of W in (7), they derive the optimal change of drift $-\nabla V$ (which corresponds to an additional biasing potential V) for the second SDE that leads to a zero-variance estimator for the quantity $F(x)$ in (7). The optimal change of drift is related to the solution of a nonlinear partial differential equation called the 'Hamilton-Jacobi-Bellman' (HJB) equation. Since the HJB equation is difficult to solve - especially in high dimensions - the authors project the bias potential into a space spanned by a finite number of ansatz functions. They suggest using a linear set of basis functions, as is done in [Valsson and Parrinello, 2014]. The problem can be reformulated as a constrained optimization problem which can be solved by a stochastic gradient descent method or by a cross entropy method [Zhang et al., 2014].

In contrast to the method described in [Valsson and Parrinello, 2014], the method in [Hartmann and Schütte, 2012] explicitly interprets the change of the potential as a change of measure. This fact is used to derive the objective function of the optimization problem (namely, the right-hand side of (21) below). Furthermore, if one considers estimators of the form (7), then the Kullback-Leibler divergence arises naturally in the derivation of the optimization problem. See Section 4.3 below for more details on the Kullback-Leibler divergence.

The main difference between the variational method of Valsson and Parrinello [Valsson and Parrinello, 2014] and the method of Hartmann and Schütte [Hartmann and Schütte, 2012] is that the first method works directly on the potential itself, while the second method works on the force field of the system. However, since the force field is the gradient the potential, and since the gradient of a function is uniquely determined up to additive constants, it follows that both methods approach the problem essentially the same way, i.e. by systematically finding the optimal importance sampling change of measure (bias). We note here that the task of designing an importance sampling bias is not straightforward and highly depends on the considered problem c.f. [Spiliopoulos, 2015, Dupuis et al., 2015]. We also note that there exist examples in which importance sampling can also lead to a speed-up in the enhanced dynamics [Hartmann et al., 2016].

The key idea of the method in [Hartmann and Schütte, 2012] is that there exists a representation formula for the Radon Nikodym derivative. The formula is given by Girsanov's theorem from stochastic analysis, which one can interpret as follows: if one changes the drift term of the SDE (13) by adding u_t to obtain (14), then under the conditions (15) and (16), the laws of the solutions to these SDEs (viewed as probability measures on path space) are mutually equivalent, and Girsanov's formula (18) describes the associated Radon-Nikodym derivative.

Theorem 1 (Girsanov's theorem) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $(B_t)_{t\geq 0}$ be the standard $\mathbb{R}^n$-valued Brownian motion with respect to $\mathbb{P}$. Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^n$ be Itô diffusions of the form*

$$dX_t = b(X_t)dt + \sqrt{2\beta^{-1}}dB_t, \qquad X_0 = x \tag{13}$$

$$dY_t = (u_t + b(Y_t))dt + \sqrt{2\beta^{-1}}dB_t, \qquad Y_0 = y \tag{14}$$

considered over the time interval $0 \leq t \leq \tau$ for some finite τ, where $b : \mathbb{R}^n \to \mathbb{R}^n$ satisfies Lipschitz and growth conditions that suffice for uniqueness and existence of the solutions. If the process u_t satisfies

$$\mathbb{P}\left(\int_0^\tau |u_t|^2 \, dt < \infty\right) = 1, \tag{15}$$

$$\mathbb{E}\left[\exp\left(\sqrt{\frac{\beta}{2}}\int_0^\tau u_t dB_t - \frac{\beta}{4}\int_0^\tau |u_t|^2 dt\right)\right] = 1 \tag{16}$$

then the path measures $\mu = \mathbb{P} \circ Y^{-1}$ and $\nu = \mathbb{P} \circ X^{-1}$ are equivalent, where the Radon-Nikodym derivative

$$\frac{d\mu}{d\nu} = M = (M_t)_{0\leq t\leq \tau} \tag{17}$$

is a stochastic process defined by

$$M_t = \exp\left(\sqrt{\frac{\beta}{2}}\int_0^t u_s dB_s - \frac{\beta}{4}\int_0^t |u_s|^2\, ds\right). \tag{18}$$

Proof *See [Robert Liptser and Shiryaev, 2001, Theorem 7.2].*

Note that Girsanov's theorem is often stated with (15) replaced by the stronger Novikov condition

$$\mathbb{E}\left[\exp\left(\frac{\beta}{4}\int_0^\tau |u_t|^2\, dt\right)\right] < \infty,$$

see e.g. [Øksendal, 2003].

In [Hartmann and Schütte, 2012], the drift term b in (13) is given by the force field $-\nabla U$ associated to the energy function U of a molecule. Thus the metastability of the molecule is determined by U. The Hartmann-Schütte method suggests adding suitably scaled Gaussians to 'fill' the basins in the energy landscape that are associated to metastable conformations. The resulting change in the potential is given by taking the sum of all the added Gaussians. This is idea is essentially identical to the approach taken in Metadynamics [Laio and Parrinello, 2002] where the potential is changed by Gaussian and the force field is changed by the derivative of the Gaussians.

Several different aspects of the Hartmann-Schütte approach have been studied. The question of whether Gaussians are a good choice of ansatz function has been considered in [Zhang et al., 2014]. A method for placing such ansatz functions automatically is presented in [Quer et al., 2017]. The convergence of the stochastic gradient descent approach has been analyzed in [Lie et al., 2015, Lie, 2016].

4.2 Application to Brownian Dynamics

We now consider Girsanov's theorem in the context of the Brownian dynamics model from MD. Let $U : \mathbb{R}^n \to \mathbb{R}$ be a metastable potential satisfying the criteria mentioned earlier. Setting $b = -\nabla U$ in (13) yields the original SDE (2) that defines the molecular dynamics in equilibrium. Now recall that in MD, one often seeks to reduce the metastability of the system, in order to sample rare events better. For example, in the ideal case in which one could optimise the algorithm parameters at every step in the MD simulation, the optimally biased potential that results from Metadynamics is flat. This ideal case can be represented in (14) by choosing $u_t = \nabla U(Y_t)$ for every t, which causes the drift term in (14) to be identically zero. In particular, this implies that the resulting stochastic process Y in is Brownian motion scaled by $\sqrt{2\beta^{-1}}$, and thus exhibits no metastability at all. From the mathematical point of view, this situation is ideal, because the Brownian motion process is well understood. By (17), sample statistics derived from trajectories of Y can be reweighted using the associated values of the Radon-Nikodym derivative given by (18) (with $u_s = \nabla U(Y_s)$) in order to obtain unbiased estimators of statistical quantities involving X.

We close this section with an observation regarding Girsanov's theorem. In stochastic analysis, Girsanov's theorem is sometimes formulated as follows: there exists an alternative probability measure $\mathbb{P}'$ that differs from the original probability measure $\mathbb{P}$ of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that the law of the

process Y with respect to $\mathbb{P}'$, $\mathbb{P}' \circ Y^{-1}$ and the law of the process X with respect to $\mathbb{P}$, $\mu = \mathbb{P} \circ X^{-1}$, agree. Equivalently, the alternative probability measure $\mathbb{P}'$ is such that

$$\hat{B}_t := \sqrt{\beta} \int_0^t u_s ds + B_t \tag{19}$$

is a Brownian motion with respect to the measure $\mathbb{P}'$, so that

$$dY_t = -\nabla U(Y_t)dt + \sqrt{2\beta^{-1}} d\hat{B}_t \tag{20}$$

is structurally identical to (13). Thus, we can calculate the quantity of interest according to

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}[f(X_{0:\tau})] &= \mathbb{E}_{\mathbb{P}'}[f(Y_{0:\tau})] \\ &= \mathbb{E}_{\mathbb{P}}[f(Y_{0:\tau})M_t]. \end{aligned}$$

4.3 Optimization and Kullback-Leibler

Both the Valsson-Parrinello and Hartmann-Schütte methods use an optimization problem to minimize the difference between a reference distribution and the auxiliary distribution. One way of quantifying the difference of two measures is the Kullback-Leibler divergence (KL) or relative entropy. Given two probability distributions μ and ν, the KL divergence of the former with respect to the latter is given by

$$H(\nu\|\mu) = \int \log \frac{d\nu}{d\mu} d\nu$$

The KL divergence is an information-theoretical tool that quantifies the difference between two probability measures. However, the KL divergence does not define a metric on the space of probability measures; it is not in general symmetric, and does not satisfy the triangle inequality.

If (17) holds, then

$$H(\nu\|\mu) = -\mathbb{E}_\nu \left[\log M_\tau\right] = \mathbb{E}_{\mathbb{P}} \left[\frac{1}{2} \int_0^\tau |u_t|^2 dt\right],$$

where we used that the stochastic integral $\int_0^\tau u_s dB_s$ has zero mean with respect to $\mathbb{P}$ because B is a Brownian motion process with respect to $\mathbb{P}$. Therefore, for the quantity $F(x)$ defined in (7), Girsanov's formula (17) and Jensen's inequality yield

$$\begin{aligned} F(x) &= -\beta \log \mathbb{E}_\mu[\exp(-W/\beta)] \\ &= -\beta \log \mathbb{E}_\nu \left[\exp\left(-W/\beta\right) \exp\left(\log M_\tau\right)\right] \\ &\leq \mathbb{E}_\nu[W] + \beta H(\nu\|\mu), \end{aligned} \tag{21}$$

In [Valsson et al., 2016, Section 9], it is described how the optimization problem in the Valsson-Parrinello method for finding the optimal biasing potential involves minimizing a Kullback-Leibler divergence.

For the Hartmann-Schütte method, the optimization problem seeks the optimal bias that leads to a zero-variance estimator of the quantity (7). In

[Hartmann and Schütte, 2012] the authors prove that such an optimal bias exists in the space of admissible control processes; note that in practice, the approximations required for numerical computations imply that the computed estimator will not yield a zero-variance estimator. The Hartmann-Schütte method involves the problem of optimizing the right-hand side of (21) over a suitable class of auxiliary distributions. In [Hartmann and Schütte, 2012] it is shown that such an auxiliary distribution exists, and that moreover the optimal ν yields an equality in (21); equality implies that the resulting estimator of $F(x)$ has zero variance. In [Zhang et al., 2014], it is shown how this optimization problem can be reformulated as the problem of minimizing a KL divergence. Thus one can apply algorithms for minimizing the Kullback-Leibler divergence, such as the cross entropy method [Rubinstein and Kroese, 2004]. The approach described in [Hartmann and Schütte, 2012] involves searching for the best possible auxiliary distribution $\mathbb{Q}$ in a class of auxiliary distributions parametrized by finite-dimensional vectors, where the vectors contain the expansion parameters of a feedback control function for a given finite collection of basis functions. Stochastic gradient descent is used to find the optimal distribution from this class.

4.4 Reweighting

To get the right statistics for the variational approach to enhanced sampling the authors suggest a reweighting scheme proposed by [Tiwary and Parrinello, 2013] which was proposed in [Valsson et al., 2016]. The authors derive a correction formula for the transition time between two basins sampled by a Metadynamics approach. They first assume that there exits collective variables $\lambda(R)$ such that $\lambda \leq \lambda^\star$ for the starting basin and $\lambda > \lambda^\star$ and $\lambda^\star$ is the transition region. Then they define the mean transition time

$$\tau = \frac{Z_0}{\omega \kappa Z_0^\star} = \frac{1 \int_{\lambda \leq \lambda^\star} e^{-\beta U(R)} dR}{\omega \kappa \int_{\lambda = \lambda^\star} e^{-\beta U(R)} dR}. \tag{22}$$

the mean transition time $\tau_M(t)$ for a Metadynamics simulation is defined by

$$\tau_M(t) = \frac{1 Z_M(t)}{\omega \kappa_M Z_M^\star(t)} \tag{23}$$

where κ_M, Z_M and $Z_M^\star$ are analogous to κ, Z_0 and $Z_0^\star$ except that the potential includes a time dependent Metadynamics potential. If the transition region is not biased meaning that $\kappa \approx \kappa_M$ and $Z_M^\star \approx Z_0^\star$ then one can define the acceleration factor $\alpha = \frac{\tau}{\tau_M(t)}$

$$\alpha(t) \approx \frac{Z_0}{Z_M} = \int e^{\beta V(s(R),t)} dR \tag{24}$$

where dR is the measure corresponding to the time-dependent Metadynamics potential. In order to use this correction formula, one must avoid biasing the potential in the transition region.

For the Hartmann-Schütte method in [Hartmann and Schütte, 2012], the Radon-Nikodym derivative (17) given by Girsanov's formula provides the correct reweighting with which we can use statistics based on the auxiliary distribution ν in order to estimate statistics based on the reference distribution μ.

We note that one can perform reweighting with the Radon-Nikodym derivative regardless of whether the importance sampling bias is optimal, see for example [Lelièvre and Stoltz, 2016]. In particular, there exist auxiliary distributions ν which are not optimal but for which the variance of the estimator of $F(x)$ with respect to ν is smaller than the variance of the corresponding estimator with respect to μ. In contrast to the reweighting method for Metadynamics presented in [Tiwary and Parrinello, 2013], the reweighting in the Hartmann-Schütte method does not involve any heuristic argument; it is fully justified by Girsanov's theorem. Furthermore, the importance sampling estimator is always an unbiased estimator, as was shown in (21). Another advantage of the importance sampling approach is that, in contrast with the reweighting scheme of [Tiwary and Parrinello, 2013] described above, there is no a priori reason to avoid biasing the dynamics in the transition region.

4.5 Other methods

In this paragraph we comment briefly on some other enhanced sampling methods that can be viewed from the perspective of importance sampling.

ABF and Metadynamics. Adaptive biasing force techniques (ABF) e.g. [Darve and Porohille, 2001] involve changing the force field that determines the molecular dynamics, and hence implicitly involve changing the potential. We observed in Section 4 that changing the potential corresponds to a change of the measure. Thus ABF-based methods can be interpreted as an importance sampling method. The same is true for Metadynamics [Laio and Parrinello, 2002] and all other biasing techniques which change the potential or the force field, e.g. [Huber et al., 1994].

Simulated Annealing. In simulated annealing [Kirkpatrick et al., 1983], the temperature of the simulated system is changed. When the dynamics are given by the overdamped Langevin equation (2), changing the temperature corresponds to changing the value of the inverse temperature parameter β. This again leads to a change of measure because the temperature is also included in the Boltzmann Gibbs measure. Thus, one can relate the two different measures by a Radon Nikodym derivative. From a theoretical point of view, one could built an importance sampling scheme with different temperatures involved.

It may be possible to construct a Multilevel Monte Carlo (MLMC) scheme from simulated annealing- or parallel tempering-based methods, by setting the levels of the estimator to be the different temperatures. Other methods, such as those presented in [Roe et al., 2014] or [Quer and Weber,] that work by changing the potential of the system, could also be used to build an MLMC estimator. For further information on MLMC methods see e.g. [Giles, 2015] and the references therein.

Replica Exchange. In the case of Replica Exchange methods
e.g. [Swendsen and Wang, 1986] the situations is more complex. If for example a temperature replica exchange is applied, then this can be interpreted as a multistage simulated annealing, and in principle one could compute the reweighting factors for each trajectory segments. However, one has to track the interchange

of the trajectories very precisely in order to correctly reweight the trajectory segments. To the best of our knowledge, no studies have determined whether this strategy is feasible.

5 Connections to other topics

We briefly mention some interesting connections between importance sampling and other active research topics. The references cited in this section are not intended to be exhaustive.

Optimal control

In Section 4.2, we described how finding the optimal change of measure in terms of variance reduction is related to the question of finding the optimal change of drift u_t in (14). The change of drift can be interpreted as a control force acting on the SDE, and so the problem can be interpreted as an optimal control problem. It is possible to derive a HJB equation for this problem [Fleming and Soner, 2006, Chapter 3, p. 119]. In [Hartmann and Schütte, 2012] it is shown that the solution of the HJB equation can be used to obtain the optimal change of drift (and hence the optimal auxiliary distribution). The connection between the HJB equation and importance sampling or variance reduction has been extensively studied; see [Dupuis and Wang, 2004, Dupuis et al., 2007, Dupuis et al., 2015, Vanden-Eijden and Weare, 2012, Hartmann and Schütte, 2012] and the references therein.

The HJB equation is a nonlinear partial differential equation, and hence is not easy to solve. Some methods have been proposed for solving the HJB equation, based on dynamical principles [Bertsekas, 2017] or viscosity solutions [Fleming and Soner, 2006]. Other approaches construct sub-optimal but effective importance sampling schemes and study how the variance is reduced; see [Spiliopoulos, 2015]. The connection between suboptimal importance sampling schemes and subsolutions of the HJB equation is explored in [Dupuis et al., 2012]. The connection between importance sampling and the HJB equation can also be studied from the perspective of large deviations. For more details on large deviations, see [Freidlin and Wentzell, 2012].

Machine Learning and Bayesian Inference

We note that Bayesian inference and machine learning techniques have been applied to solve Markov decision problems, see e.g. [Bierkens and Kappen, 2014, Thijssen and Kappen, 2015, Bertsekas, 2017]. These problems are similar to stochastic optimal control problems in the sense that they admit a HJB equation. An excellent exposition of the connection between Bayesian inference, machine learning, and Markov decision processes is given in [Kappen, 2013]. Kernel estimation methods have also been applied in this context [Batz et al., 2016, Opper, 2017].

6 Summary

In this article, we set molecular dynamics and enhanced sampling in the framework of mathematical probability theory. We described the basic idea of importance sampling and showed that enhanced sampling can be viewed as importance sampling on path space. We demonstrated this observation using the recently developed Valsson-Parrinello variational enhanced sampling method [Valsson and Parrinello, 2014]. We compared the Valsson-Parrinello method with the Hartmann-Schütte method for efficient simulation of rare events [Hartmann and Schütte, 2012], and described how both methods involve 1) a parametrization of the bias potential in order to set up an optimization problem involving the Kullback-Leibler divergence, and 2) a formula for reweighting statistics obtained from a biased auxiliary distribution, in order to obtain statistics from a desired reference distribution. We mentioned how other enhanced sampling methods, such as the adaptive biasing force method, can be viewed as a method for importance sampling. The fundamental criteria that justifies viewing an enhanced sampling method as an importance sampling method is 1) that a change of measure is realised, e.g. by a change in the potential energy function (or equivalently, of the force field) that describes the molecular dynamics, and 2) that the empirical biased statistics can be reweighted to obtain unbiased estimators of the true statistics.

Acknowledgements

The research of JQ is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft through grant CRC 1114 "Scaling Cascades in Complex Systems", Project A05 'Probing scales in equilibrated systems by optimal nonequilibrium forcing'. HCL is supported by the Free University of Berlin within the Excellence Initiative of the German Research Foundation.

References

[Batz et al., 2016] Batz, P., Ruttor, A., and Opper, M. (2016). Variational estimation of the drift for stochastic differential equations from the empirical density. *J. of Stat. Mech. Theo. and Exper.*, 2016(8):083404.

[Bernardi et al., 2015] Bernardi, R., Melo, M., and Schulten, K. (2015). Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Bio. et Biophy. Acta (BBA) - General Subjects*, 1850(5):872 – 877. Recent developments of molecular dynamics.

[Bertsekas, 2017] Bertsekas, D. P. (2017). *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, 4 edition.

[Bierkens and Kappen, 2014] Bierkens, J. and Kappen, H. J. (2014). Explicit solution of relative entropy weighted control. *Syst. & Cont. Lett.*, 72:36–43.

[Bucklew, 2004] Bucklew, J. (2004). *Introduction to Rare Event Sampling*. Springer Series in Statistics. Springer-Verlag New York, New York.

[Cerou and Guyader, 2006] Cerou, F. and Guyader, A. (2006). Adaptive multilevel splitting for rare event analysis. *Stoch. Anal. and Appl.*, 25(2).

[Chipot and Pohorille, 2007] Chipot, C. and Pohorille, A. (2007). *Free Energy Calculations*. Springer Series in Chemical Physics. Springer, Berlin Heidelberg.

[Darve and Porohille, 2001] Darve, E. and Porohille, A. (2001). Calculating free energy using average forces. *J. for Chem. Phys.*, 115:9169–9183.

[Dupuis et al., 2007] Dupuis, P., Sezer, A. D., and Wang, H. (2007). Dynamic importance sampling for queueing networks. *A of Appl. Prob.*, 17(4):1306–1346.

[Dupuis et al., 2015] Dupuis, P., Spiliopoulos, K., and Zhou, X. (2015). Escaping from an attractor: Importance sampling and rest points i. *A. of Appl. Prob.*, 25(5):2909–2958.

[Dupuis and Wang, 2004] Dupuis, P. and Wang, H. (2004). Importance sampling, large deviations, and differential games. *Stoch. and Stoch. Rep.*, 76(6):481–508.

[Dupuis et al., 2012] Dupuis, P., Wang, H., and Spiliopoulos, K. (2012). Importance sampling for multiscale diffusions. *SIAM Multis. Model. and Sim.*, 12:1–27.

[Elber and West, 2010] Elber, R. and West, A. (2010). Atomically detailed simulation of the recovery stroke in myosin by milestoning. *PNAS*, 107(11).

[Fleming and Soner, 2006] Fleming, W. and Soner, H. (2006). *Controlled Markov Processes and Viscosity Solutions*. Springer-Verlag New York, New York, 2 edition.

[Freidlin and Wentzell, 2012] Freidlin and Wentzell (2012). *Random Perturbations to Dynamical Systems*. Springer-Verlag Berlin Heidelberg, Berlin, 3 edition.

[Giles, 2015] Giles, M. (2015). Multilevel monte carlo methods. *Acta Numerica*, 24:259–328.

[Hartmann and Schütte, 2012] Hartmann, C. and Schütte, C. (2012). Efficient rare event simulation by optimal nonequilibrium forcing. *J Stat. Mech. Theo. and Exper.*, 2012.

[Hartmann et al., 2016] Hartmann, C., Schtte, C., and Zhang, W. (2016). Model reduction algorithms for optimal control and importance sampling of diffusions. *Nonlinearity*, 29(8):2298.

[Huber et al., 1994] Huber, T., Torda, A. E., and van Gunsteren, W. F. (1994). Local elevation: A method for improving the searching properties of molecular dynamics simulation. *J. of Comp.-Aid. Mol. Des.*, 8:695–708.

[Kappen, 2013] Kappen, H. J. (2013). Talk. `http://www.snn.ru.nl/~ber/acns/summerschool_nijmegen2013.pdf`.

[Kirkpatrick et al., 1983] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

[Laio and Parrinello, 2002] Laio, A. and Parrinello, M. (2002). Escaping free-energy minima. *PNAS*, 20(10):12562–12566.

[Leimkuhler and Matthews, 2015] Leimkuhler, B. and Matthews, C. (2015). *Molecular Dynamics*. Springer International Publishing.

[Leliévre et al., 2010] Leliévre, T., Rousset, M., and Stoltz, G. (2010). *Free energy computations : a mathematical perspective*. Imperial College Press, London, Hackensack (N.J.), Singapore.

[Lelièvre and Stoltz, 2016] Lelièvre, T. and Stoltz, G. (2016). Partial differential equations and stochastic methods in molecular dynamics. *Acta Numerica*, 25:681–880.

[Lie et al., 2015] Lie, H., Schütte, C., and Hartmann, C. (2015). Martingale-based gradient descent algorithm for estimating free energy values of diffusions. `http://publications.imp.fu-berlin.de/1476/1/MGD_03b.pdf`.

[Lie, 2016] Lie, H. C. (2016). Convexity of a stochastic control functional related to importance sampling of It diffusions. arxiv.org/abs/1603.05900.

[Øksendal, 2003] Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*. Hochschultext / Universitext. Springer, Berlin.

[Opper, 2017] Opper, M. (2017). An estimator for the relative entropy rate of path measures for stochastic differential equations. *J. of Comp. Phys.*, 330:127 – 133.

[Quer et al., 2017] Quer, J., Donati, L., Keller, K., and Weber, M. (2017). An automatic adaptive importance sampling algorithm for molecular dynamics in reaction coordinates. *SAIM J. of Sci. Comp.* submitted.

[Quer and Weber,] Quer, J. and Weber, M. Estimating exit rates in rare event dynamical systems via extrapolation. Technical report, Zuse Institut Berlin.

[Robert Liptser and Shiryaev, 2001] Robert Liptser, R. and Shiryaev, A. (2001). *Statistics of Random Processes I. General Theory*. Stochastic Modelling and Applied Probability. Springer-Verlag Berlin Heidelberg, Berlin Heidelberg.

[Roe et al., 2014] Roe, D., Bergonzo, C., and Cheatham, T. (2014). Evaluation of enhanced sampling provided by accelerated molecular dynamics with hamiltonian replica exchange methods. *J. of Phys. Chem. B*, 118(13).

[Rubinstein and Kroese, 2004] Rubinstein, R. and Kroese, D. (2004). *The Cross-Entropy Method*. Info. Sci. and Stat. Springer-Verlag New York, New York.

[Sorensen and Voter, 2000] Sorensen, M. R. and Voter, A. F. (2000). Temperature-accelerated dynamics for simulation of infrequent events. *J. of Chem. Phys.*, 112(21):9599–9606.

[Spiliopoulos, 2015] Spiliopoulos, K. (2015). Non-asymptotic performance analysis of importance sampling schemes for small noise diffusions. *J. of Appl Prob.*, 52:1–14.

[Swendsen and Wang, 1986] Swendsen, R. and Wang, J. (1986). Replica monte carlo simulation of spin glasses. *Phy. Rev. Lett.*, 57.

[Thijssen and Kappen, 2015] Thijssen, S. and Kappen, H. J. (2015). Path integral control and state-dependent feedback. *Phys. Rev.*, 91(3).

[Tiwary and Parrinello, 2013] Tiwary, P. and Parrinello, M. (2013). From metadynamics to dynamics. *Phys. Rev. Lett.*, 111:230602.

[Torrie and Valleau, 1977] Torrie, G. and Valleau, J. (1977). Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *J. of Comp. Phys.*, 23(2):187–199.

[Tuckerman, 2010] Tuckerman, M. (2010). *Statistical Mechanics: Theory and Molecular Simulation*. Oxford Graduate Texts. OUP Oxford.

[Valsson and Parrinello, 2014] Valsson, O. and Parrinello, M. (2014). Variational approach to enhanced sampling and free energy calculations. *Phys. Rev. Lett.*, 113(9):090601.

[Valsson et al., 2016] Valsson, O., Tiwary, P., and Parrinello, M. (2016). Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *A. Rev. of Phys. Chem.*, 67(1):159–184.

[Vanden-Eijden and Weare, 2012] Vanden-Eijden, E. and Weare, J. (2012). Rare event simulation of small noise diffusions. *Comm. on Pure and Appl. Math.*, 65:1770 –1803.

[Voter, 1997] Voter, A. F. (1997). A method for accelerating the molecular dynamics simulation of infrequent events. *J. of Chem. Phys.*, 106(11):4665–4677.

[Wang and Landau, 2001] Wang, F. and Landau, D. P. (2001). Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys. Rev. E*, 64:056101.

[Zhang et al., 2014] Zhang, W., Wang, H., Hartmann, C., Weber, M., and Schuette, C. (2014). Applications of the cross-entropy method to importance sampling and optimal control of diffusions. *J. of Sci. Comp.*, 36.